CSC 478 – Final Project Report
June 5, 2017
Brandon Markwalder, William Martin, David Scroggins

1. Overview of Dataset

The National Longitudinal Study of Adolescent Health tracked American adolescents grades 7-12 from 1994 to 2008. The dataset was developed from the Quality Education Database, which contains 26,666 U.S. High Schools. The complete dataset can be accessed here: http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/21600 along with relevant user guides for study design, variable description etc.

A stratified sample of 80 schools with "probability of selection proportional to school size" were selected. Schools were stratified by region, urbanicity, school type, ethnic mix and size. One feeder school for each high school was recruited producing one school pair in 80 different communities. School size varied from 100 to 3,000+ students. Most the dataset was produced via a sequence of surveys. First, an In-School Survey in 1994 completed by over 90,000 students. The Wave 1 In-Home Survey (1995), in which adolescents were selected with unequal probability of selection from those who complete in-school survey was drawn from the In-School survey via stratifications by grade and sex and then random selection of ~17 students from each stratum to yield a sample of 200 adolescents from each school pair. The In-Home survey was repeated three additional times: Wave 2 In-Home Survey (1996); Wave 3 In-Home Survey (2001) and Wave 4 In-Home Survey (2008) (sampling criteria for each wave is detailed in the study's downloadable User Guide). In addition to survey results, Wave 4 includes several biomarker datasets including inflammation and immune function, measures of glucose homeostasis as well as lipids.

The composition of the dataset varies by Wave, however, the results from the "Wave 1: In-Home Questionnaire" give an adequate sense of scale. The dataset contains 2,794 columns/variables and 6,281 rows/instances for a total of 17,549,114 unique data points. A thorough summary of the variables is beyond the scope of this proposal. The variables range from basic demographic questions: date of birth, racial and ethnic background, intimate partner status, school status, daily activities, general health, access to health services, parental relationships, pregnancy history, relations with siblings etc. They provide a comprehensive self-reported description of the student's current living situation and personality. Complete details can be obtained from the user guides available at the dataset's main site.

2. Description of the Problem

The complexity of the dataset presents some significant challenges. Given the relatively comprehensive nature of the data, a multitude of interesting questions and suitable techniques present themselves. We began with a careful review of the documentation, looking for potential patterns and interesting questions that were answerable in a machine learning context.

After initial review, a consensus emerged that the Wave 4 Biomarkers, which contain a thorough array of testing for glucose homeostasis, lipids and measures of immune function provided a clearly articulated target for analysis. The decision was made to focus on these biomarkers as our classes and explore the dataset to answer the general question of whether we could use the subjective reporting of the In-Home Questionnaires to generate a composite image of the social structure surrounding individuals long term health as objectively indicated in the Wave 4 biomarkers. The question was not causal, as in many cases we cannot claim that certain factors directly cause a given condition. This is particularly true in the case of classifying diabetics given Type 1 and Type 2 diabetes have different and complex connections to socioeconomic status, race, genetics as well as dietary habits.

The decision was made to focus initially on the Wave 1 In-Home Questionnaire. This presents a risk in that the responses were given a full 14 years prior to the collection of the biomarkers. However, it also presents an exciting opportunity to explore the association between early self-reporting and long term health. Furthermore, it provides a ground floor for analyzing the ways in which individuals' social dynamics change over time and consequently serve as predictors for long term health. The intent was to generate as comprehensive of an analysis of Wave 1 followed by an initial exploration of the classification potential of the Wave 4 In-Home Questionnaire, which is synchronous with biomarker collection. The long-term plan, which goes beyond the scope of this class, is to generate a complete analysis of all four waves' association with Wave 4 biomarkers as well as a granular analysis of the ways in which individuals self-reporting changes over time.

3. Initial Dataset Survey

Wave 1 In-Home Questionnaire contains 2,794 features (responses) and 6,281 instances (students). The data are complete, however, the structure of the questionnaire, with many questions being conditionally answered based on whether a student has answered a previous question, produced a noisy dataset with many data points contain special encoding indicating a question was skipped. Since this encoding in general does not contribute to meaningful interpretation, careful initial selection of a subset of features was essential. We proceeded with an initial inspection of the Codebook for the Wave 1 In-Home Questionnaire to identify, based on domain knowledge, which questions fit a general criterion of a 'clearly articulated social or behavioral factor whose association with long term health can be meaningfully interpreted.' As an example of a question that satisfies this question consider H1WP8, which asks, "On how many of the past 7 days was at least one of your parents in the room with you while you ate your evening meal?" [1] Such a question presents both an objective quantitative metric for evaluation as well as a clear interpretation should it prove predictive given the widely-recognized link between dietary habits and long term health. As an example of a

[1] For easy access to the questions asked, ICRSPR provides a search feature for specific variables: http://www.icpsr.umich.edu/icpsrweb/ICPSR/ssvd/studies/21600/variables?q=h1wp8. Additional information regarding encoding of responses as well as frequencies can be located in the Codebook for each relevant wave.

question that does not satisfy this question consider H1ID5 which asks, “Have you taken a public or written pledge to remain a virgin until marriage?” While such a question presents interesting context for an exploration of pregnancy risk, it does not provide an easily interpretable association with long term health. Should it prove predictive of, for example, elevated lipid levels, it is not clear how this could potentially serve as a mechanism for long term health risks. We selected 502 features based on this criterion, links to summary of our initial report (along with scripts to replicate our analysis) can be found in the appendix to this report.

4. Dataset Preparation

Complete Python scripts for the preparation of the data set can be viewed in the attached zip file as well as downloaded from the repositories linked in the appendix. All four waves were initially prepared to allow for easier expansion of analysis. What follows is a general overview of the process. The files for each In-Home questionnaire (i.e. Waves I, II, III and IV) and the three biomarker files were imported and inspected for NaNs. The Public Use Contextual Database files (which contain the socioeconomic data for each student in Waves I and II) were imported as well. Waves I and II questionnaires were merged with the contextual database on the shared AID field, which contains a unique identifier for each student.[2] All four Waves were then merged with the biomarker datasets on the AID field.

For the Wave 1 merged filed we then filtered the dataset to include only features we manually selected in the review process. Biomarker features irrelevant to classification were dropped (for example C_HBA1C which was used to generate the classification variable C_JOINT was dropped given it contained redundant information captured in C_JOINT). Remaining NaNs were dropped (7 instances in total) to produce three datasets containing relevant selected features and the biomarker classifiers (explained below).

In the final step, dummy variables were generated for categorical features in the cleaned dataset. For Wave 1, 106 features were transformed into dummy variables. The resulting dummy variables were then filtered to remove meaningless dummy variables.[3]

5. Class Overview

Each of 3 datasets (per Wave) includes a classifier variable for the biomarker dataset as outlined below:

1. Measure of Inflammation and Immune Function: C_CRP
    1.1. *High Sensitivity C-Reactive Protein (hsCRP):* “CRP is produced by the liver in response to inflammation. It also is a fairly stable protein that can be sensitively measured with

[2] Datasets were merged on this field, since we were only interested in students who responded to the In-Home Questionnaires *and* were tested in the Wave 4 biomarkers.

[3] For instance, respondents who legitimately skipped a question based on a previous answer. In some cases, these legitimate skips were retained as they capture valuable information, for instance, H1RM1, where a legitimate skip indicates that the respondent has no biological mother.

precision using standardized laboratory procedures (Pearson et al, 2003). Moreover, in asymptomatic, intermediate-risk men aged ≤ 50 years and women ≤ 60 years, measurement of hsCRP may be useful in cardiovascular risk assessment (Greenland et al, 2010)."[4]

1.2. *C_CRP:* captures hsCRP test results defined to approximate tertiles in adult population. It is discretized into three bins:
   1.2.1. < 1 : Low
   1.2.2. 1-3 : Average
   1.2.3. > 3 : High

2. Measures of Glucose Homeostasis: C_JOINT
   2.1. *C_FGLU:* "The classification of glucose concentrations among Add Health respondents who were fasting (≥ 8 hr) at the time of blood collection was constructed based on the 2011 clinical practice recommendations for the diagnosis and classification of diabetes (ADA, 2011). "[5]
   2.2. *C_NFGLU:* "The classification of glucose concentrations among Add Health respondents who were non- fasting (< 8 hr) at the time of blood collection was constructed based on the 2011 clinical practice recommendations for the diagnosis and classification of diabetes (ADA, 2011)."[6]
   2.3. *HBA1C*: "HbA1c was assayed ... because it is an integrated measure of glucose homeostasis, reflecting average blood glucose over the preceding two to three months. The measure plays a critical role in the management of diabetes since it is correlated with micro- and macrovascular complications and is widely used as the standard biomarker for the adequacy of glycemic management."[7]
   2.4. *C_MED:* "Respondents who report using medication in the past four weeks associated with one or more of the following therapeutic classification codes:"[8]
   2.5. *C_JOINT:* an engineered feature to classify individuals as diabetic based on the following criteria:
      2.5.1. Fasting glucose >= 126 mg/dl (C_FGLU = 3)
      2.5.2. Or Non-fasting glucose >= 200 mg/dl (C_NFGLU = 2)
      2.5.3. Or HbA1c >= 6.5% (C_HBA1C = 3)
      2.5.4. Or self-reported history of diabetes (H4ID5D = 1)
         2.5.4.1. "Has a doctor, nurse or other health care provider ever told you that you have or had: high blood sugar or diabetes {if female add, when you were not pregnant}? "
         2.5.4.2. Wave 4: In-Home Questionnaire
      2.5.5. Or used anti-diabetic medication in past four weeks (C_MED = 1)

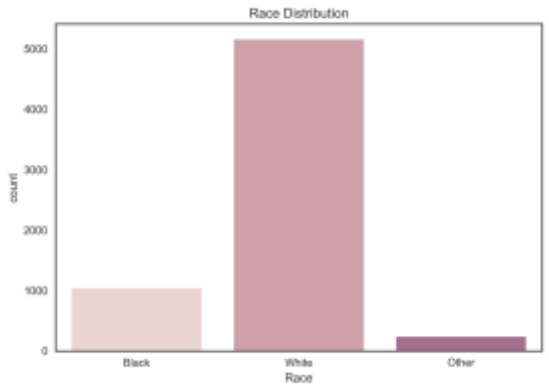
3. Lipids: C_JOINT2

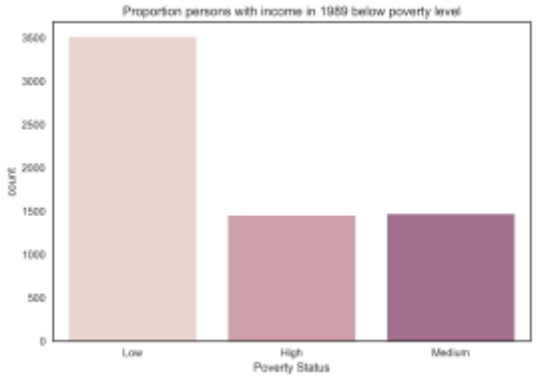
[4] "Wave 4: Biomarkers, Measures of Inflammation and Immune Function Report," p. 14
[5] Wave 4: Biomarkers, Measures of Glucose Homeostasis Report, p. 12.
[6] Ibid.
[7] Ibid., p. 13.
[8] Medication list and codes can be found on Ibid.

3.1. *C_MED2:* "Respondents using an antihyperlipidemic medications in the past four weeks with one of the following therapeutic classification codes:"[9]
3.2. *C_JOINT2*: an engineered feature to classify individuals as diabetic based on the following criteria:
  3.2.1. Self-reported history of hyperlipidemia (H4ID5B =1)
    3.2.1.1. "Has a doctor, nurse or other health care provider ever told you that you have or had: high blood cholesterol or triglycerides or lipids?"
    3.2.1.2. Wave 4: In-Home Questionnaire
  3.2.2. Used an antihyperlipidemic medication in past four weeks (C_MED2=1)

6. Overview of Dataset Visualizations

There are too many dataset features to adequately visualize here, but the following gives a general sense of the social and racial distribution of the dataset to aid with interpretation. The graphics are based on the Public Contextual Database and provide information about the neighborhoods students live in and not about the students themselves.

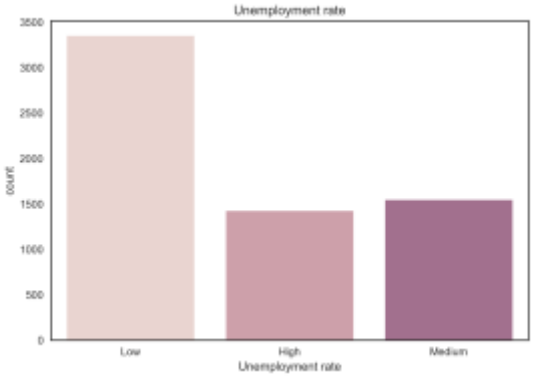


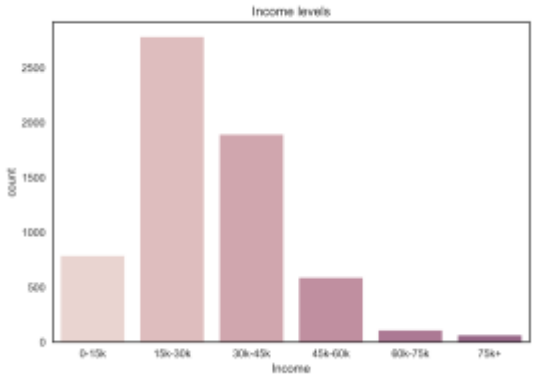


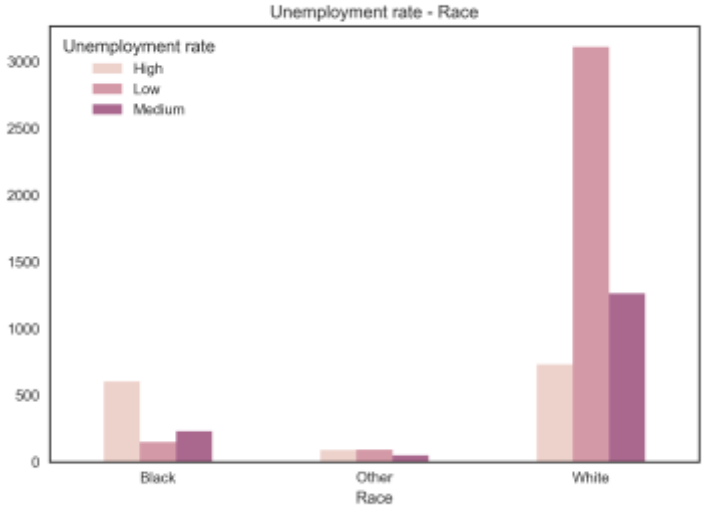
The dataset is composed heavily of white individuals, followed by black individuals and then other races. A high proportion of the individuals in the dataset are low income.

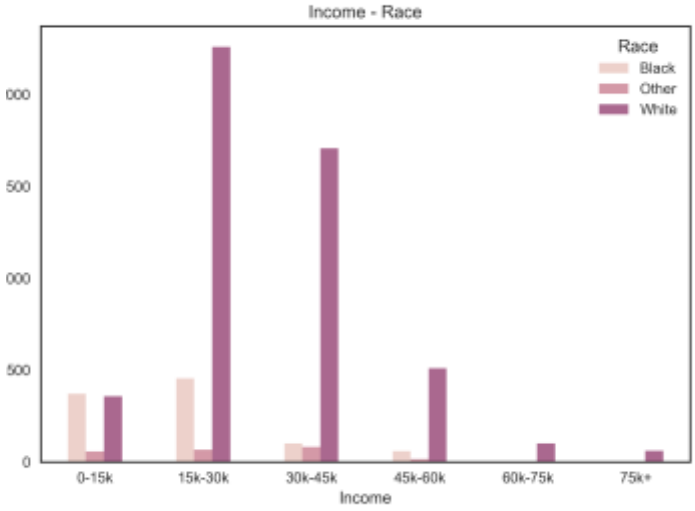
A majority if the individuals surveyed live in areas of low or medium unemployment with income level distribution skewed right (as is typically the case with income level distribution in a representative sample.

[9] Medication list and codes can be found on "Wave 4: Biomarkers, Lipids Report", p. 20

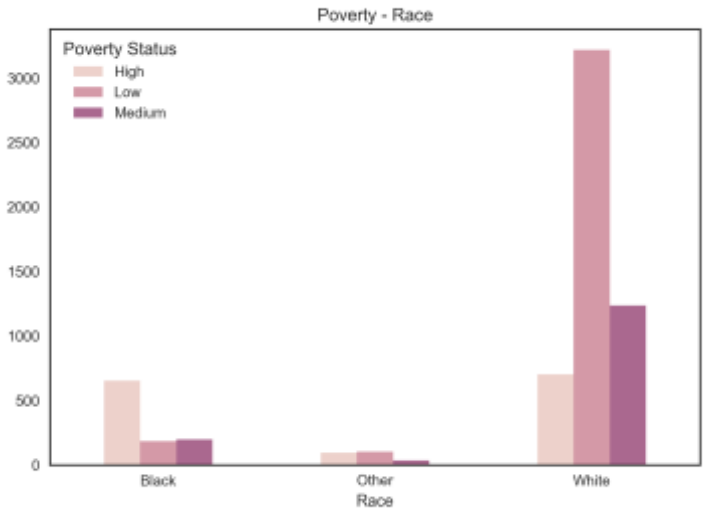


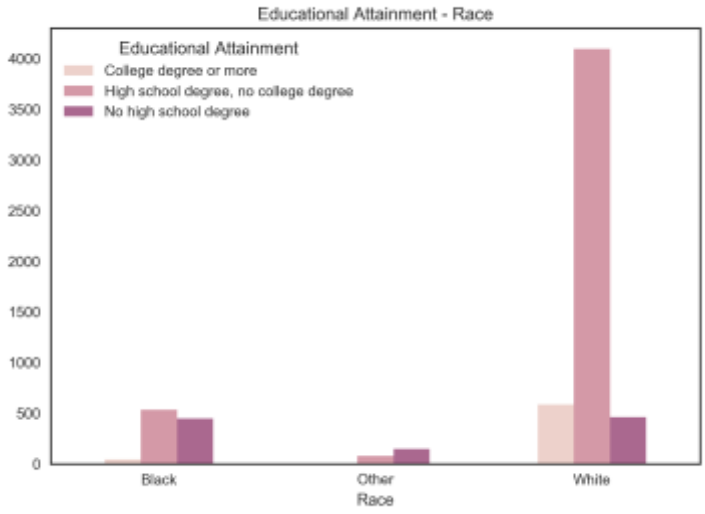






When cross-tabbed with race, the dataset clearly also coheres to other nationally representative patterns. White individuals generally enjoy a lower unemployment rate, while the unemployment rate is disproportionately high among black individuals as well as other nonwhite individuals. The higher income brackets are also disproportionately white as compared to disproportionate representation of black individuals in the lower income brackets.





These general patterns hold when cross-tabulated for poverty and educational level. Black individuals are disproportionately likely to live in poverty whereas other nonwhite individuals are also more likely to live in poverty than white individuals. Likewise, disproportionately large numbers of nonwhite individuals have no high school degree, though high school degrees are more common among black individuals than among nonwhite, nonblack individuals.

7. Initial Models

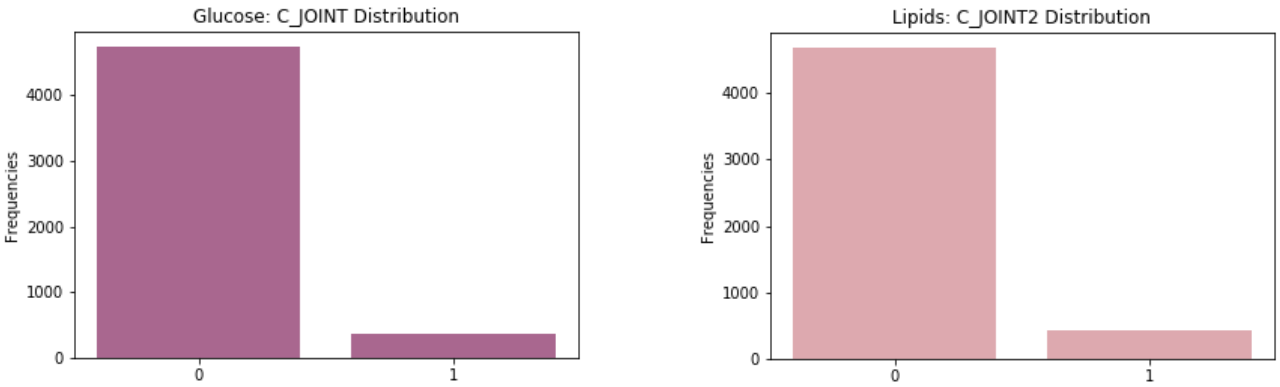
Our initial plan was to use Decision Trees both as a technique for dimensionality reduction as well as classification models in and of themselves. Decision trees are particularly well-suited for mixed, complicated data as their nonparametric approach requires no prior knowledge. They are additionally well-suited for usage with medical prediction in that they provide clear, interpretable results. Additionally, we wanted to extend the general usage of Decision Trees into the powerful ensemble learning method, Random Forest. While more difficult to interpret given their frequent 'black box' status, random forests in conjunction with decision trees promised to provide more powerful predictions that could be understood along the interpreted results of the trees. We also explored using AdaBoost to amplify our decision trees.

In general, our decision tree models exhibited a high degree of variance. We saw strong results on training data, but weak results on test data (recall often less than 10% in test). It was clear we needed to control for variance in the tree, and most of our efforts were focused on techniques for addressing this problem (weighting, oversampling, etc.). To control for variance, we experimented with several hyper-parameters. We worked primarily with tree depth limits and sample limits for leaf nodes and splitting nodes. Some experimentation with class weighting was included.

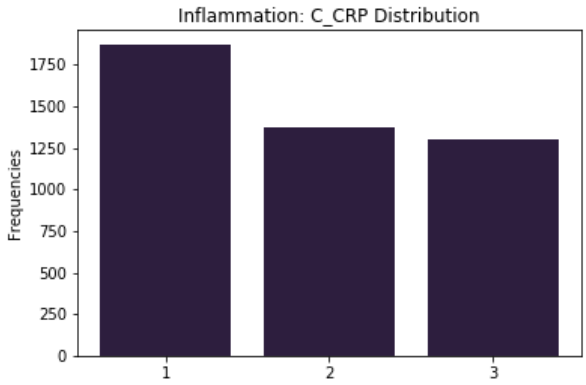
Unless otherwise noted, we used the DecisionTreeClassifier in the sklearn.tree module and RandomForestClassifier in the sklearn.ensemble module for classification. Additionally, unless otherwise stated, the assumption can be made that the dataset for each biomarker was split into training and testing sets using an 80/20 split. For initial models, the training and testing sets were also stratified based on the target feature (given the imbalance of the classes, we wanted to ensure that both the training and testing sets contained a proportional distribution of the majority and minority classes). While accuracy was considered, our analysis privileges recall on the minority class with precision on the minority class as a secondary metric. Given that the ideal is to capture information about people with 'unhealthy' biomarkers, we are most concerned with catching positive cases (recall) as well as ensuring that our positive predictions are capturing valid information (precision).

In general, the analysis for all model building in this report proceeded in the following manner: 1) An initial naïve model was trained with no attempt to tune parameters to generate a general 'ground floor' for accuracy. 2) The model parameters were then tuned using GridSearchCV from the sklearn.model_selection module with 10-fold cross validation using both accuracy and recall as scoring metrics, 3) Final models were then trained using the tuned parameters and validated using the testing split. In the case of Decision Trees, visualizations were generated of the final models.

As mentioned two of our classes were heavily imbalanced.



In the case of C_JOINT the ratio was 4746 : 361. In the case of C_JOINT2 the ratio was 4686 : 421.

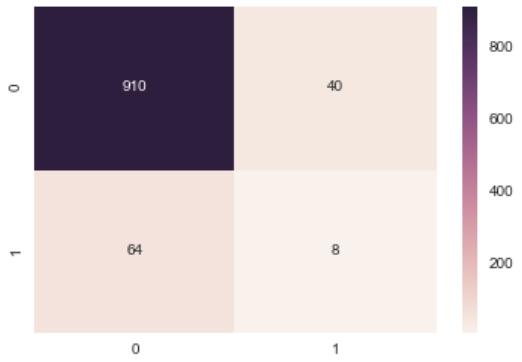


C_CRP was more evenly distributed with a ratio of 1870 : 1371 : 1298

7a. Glucose Initial Decision Tree Model

As expected, given class imbalance, a single decision tree did not perform particularly well on this dataset. While the overall accuracy for a classifier (entropy, max depth = 10, min samples in leaf = 1, min samples for split = 2) was impressive, this was due to the strong class imbalance. Recall for the minority class (C_JOINT=1) was 0.11 with a precision of 0.17.

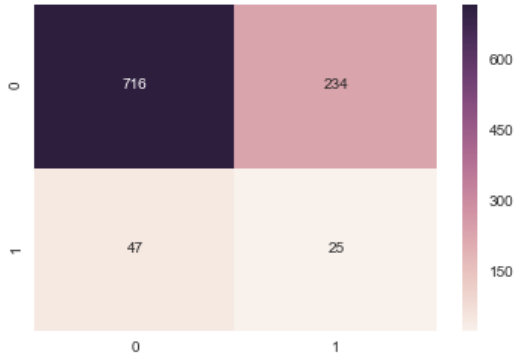
	precision	recall	f1-score	support
0	0.93	0.96	0.95	950
1	0.17	0.11	0.13	72
avg / total	0.88	0.90	0.89	1022



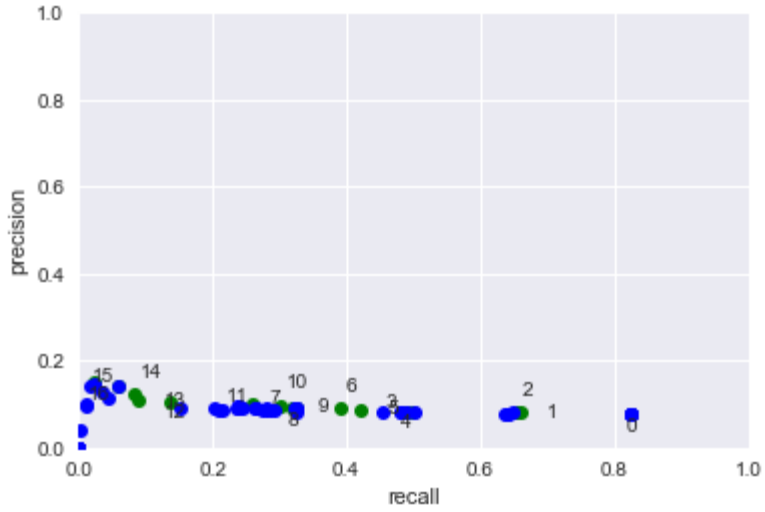
In response to this observation, we experimented with the class weight parameter to emphasize the minority class instances. Experimenting with different class weight values, we made noticeable improvements, prioritizing recall over overall precision. We eventually found ourselves at another extreme: by setting very high class weight values for C_JOINT=1 class, we

could produce perfect recall by predicting almost everyone will have diabetes (terrible precision). Reducing the extreme values in our class weights, we got to a model that performed decently in comparison to our initial tree (up to over 30% recall accuracy in test), but provided overall poor results. However, it should be noted that this model still leans heavily on weighting (10:1 weighting for minority class) which is likely indicative of overfitting (entropy, max_depth=4, min_samples_left=1, min_samples_split=2).

	precision	recall	f1-score	support
0	0.94	0.75	0.84	950
1	0.10	0.35	0.15	72
avg / total	0.88	0.73	0.79	1022



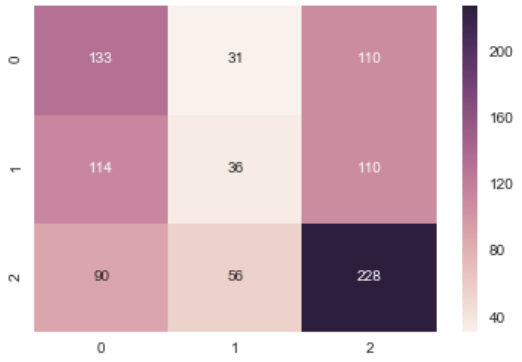
Although we still had a high number of false positives, we continued to believe in the importance of recall over precision (in moderation). We experienced a clear tradeoff between the two as we searched the hyper-parameter space, and produced a graph (of average test scores) to illustrate the tradeoff. Our above model is #10 in the graph below (green points are the best scores).



7b. Inflammation and Immune Function Initial Decision Tree Model

Immune Function differs from the other two biomarkers in that is tri-class rather than binary with instances relatively evenly distributed between the three. An initial tree (gini, max_depth=6, min_samples_leaf=11, min_samples_split=102) produced better results for the target class (C_CRP=3) with a recall score of 0.61 and precision of 0.51.

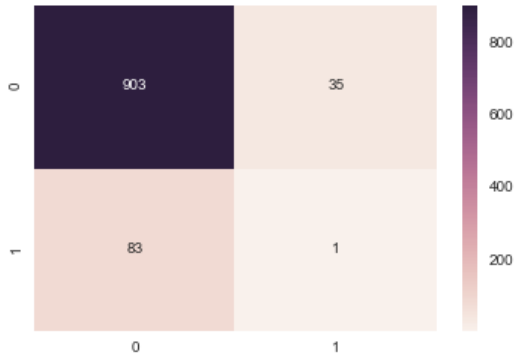
	precision	recall	f1-score	support
1.0	0.39	0.49	0.44	274
2.0	0.29	0.14	0.19	260
3.0	0.51	0.61	0.55	374
avg / total	0.41	0.44	0.41	908



7c. Lipids Initial Decision Tree Model

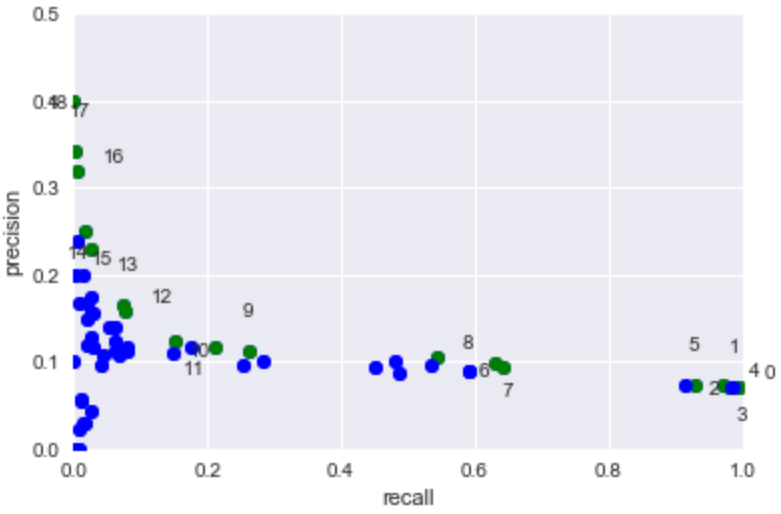
Our initial tree classifier for the lipids biomarker (C_JOINT2) suffered from the same problems as our glucose model, failing almost entirely to capture information regarding the minority class (C_JOINT2=1) with a recall score of 0.01 and precision of 0.03 (gini, max_depth=10, min_samples_leaf=1, min_samples_split=2).

	precision	recall	f1-score	support
0	0.92	0.96	0.94	938
1	0.03	0.01	0.02	84
avg / total	0.84	0.88	0.86	1022



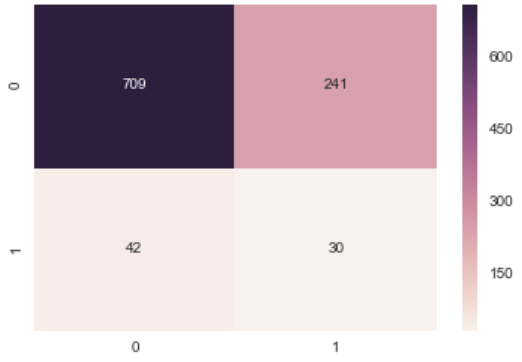
7d. Glucose Initial Random Forest

At this point, we were still aggressively experimenting with weighting, and our random forest models displayed the same tradeoff between precision and recall.



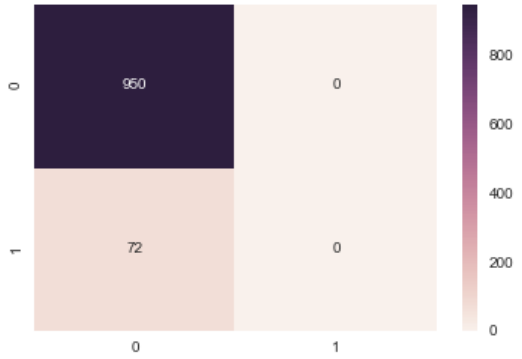
We were able to boost our recall noticeably using extremely imbalanced weighting toward the minority class, however, the heavy weighting necessary to produce the following results emphasizes the extent of the challenge (entropy, max_depth, min_samples_split=8, n_estimators=128, class_weight={0:1, 1:32}). Furthermore, even with the increase in recall emphasizes the tradeoff in precision, as only 11% of predicted minority class instances are minority class instances.

	precision	recall	f1-score	support
0	0.94	0.75	0.83	950
1	0.11	0.42	0.17	72
avg / total	0.89	0.72	0.79	1022



An unweighted random forest replicates the same classification problem as the initial decision tree (gini, max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=10).

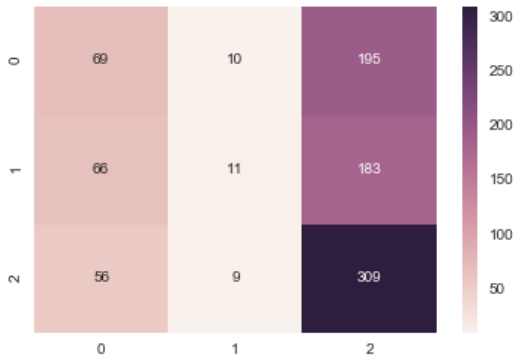
	precision	recall	f1-score	support
0	0.93	1.00	0.96	950
1	0.00	0.00	0.00	72
avg / total	0.86	0.93	0.90	1022



7e. Inflammation and Immune Function Initial Random Forest

Our initial random forest model for the inflammation and immune function performed well in terms of the target class (C_CRP=3) with a recall score of 0.83 (entropy, max_depth=7, min_samples_leaf=21, min_samples_split=43, n_estimators=50). The precision score for the target class was lower than we would have liked, however, at 0.45.

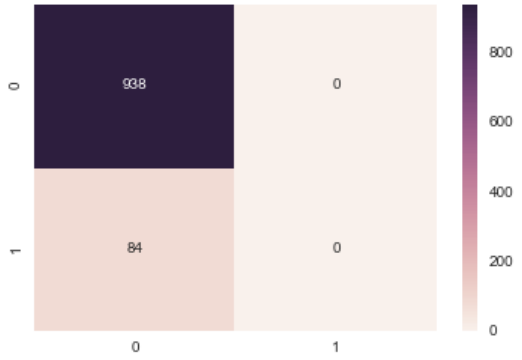
	precision	recall	f1-score	support
1.0	0.36	0.25	0.30	274
2.0	0.37	0.04	0.08	260
3.0	0.45	0.83	0.58	374
avg / total	0.40	0.43	0.35	908



7f. Lipids Initial Random Forest

An unweighted random forest performed poorly on the lipids dataset, failing to properly classify any of the minority class.

	precision	recall	f1-score	support
0	0.92	1.00	0.96	938
1	0.00	0.00	0.00	84
avg / total	0.84	0.92	0.88	1022

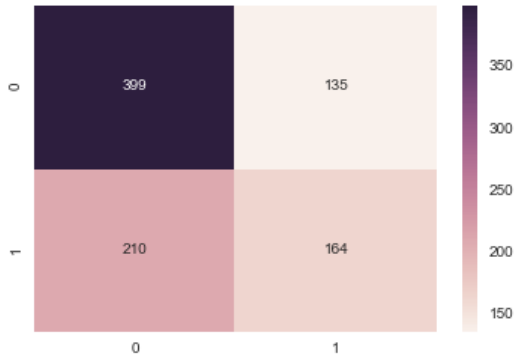


7g. Inflammation and Immune Function Initial Decision Tree (Two classes)

By way of experimentation, we decided to also consider a two-class variation of the inflammation biomarker. C_CRP=1 and C_CRP=2 ("low" and "average") were collapsed into a single class while C_CRP=3 ("high") was left as the second class. The rationale was that this might allow to more precisely target the minority class as well as apply oversampling to the dataset (see below). The initial decision tree for this set up, as expected, produced arguably worse results than the three-class model (entropy, max_depth=1, min_samples_leaf=1,

min_samples_split=2). Also of note is that GridSearchCV consistently returned these parameters, indicating that in this case, there is a single feature doing most of the classification work.

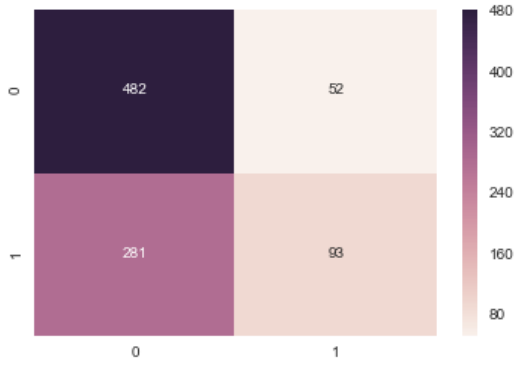
	precision	recall	f1-score	support
1.0	0.66	0.75	0.70	534
3.0	0.55	0.44	0.49	374
avg / total	0.61	0.62	0.61	908



7h. Inflammation and Immune Function Initial Random Forest (Two classes)

Random forests did not perform well on the two-class dataset, producing a recall score of 0.25 for the minority class with precision of 0.64 (entropy, max_depth=10, min_samples_leaf=1, min_samples_split=2, n_estimators=90).

	precision	recall	f1-score	support
1.0	0.63	0.90	0.74	534
3.0	0.64	0.25	0.36	374
avg / total	0.64	0.63	0.58	908



8. Oversampling

Given the problem of imbalanced classes and the insufficiency of stratified sampling and class weighting to provide more information regarding the minority classes, we explored additional options for extracting knowledge from the dataset. An additional technique that emerged as a contender was that of random oversampling in which the minority class is intentionally oversampled at random, producing duplicate instances. The purpose of the technique is to balance out the dataset to allow the models to more clearly 'see' the minority class through the noise of a large, imbalanced dataset. Some caveats apply. Because the technique intentionally alters the structure of the original dataset, models produced are not necessarily generalizable to the original dataset. Additionally, because existing instances (in this case individual students) are replicated, the danger of overfitting our models becomes a real possibility. To avoid this, it is necessary to generate cross-validation folds prior to resampling to minimize the extent to which the same instance is used twice to validate the model. Considering this, when testing the model, it is important to cross-validate the predictions rather than simply generating predictions based on the entire test set.[10]
To oversample, we utilized the RandomOverSampler module from the third-party package *imblearn* a sklearn compatible package for dealing with class imbalance.[11]

Unless otherwise stated, the oversampling technique was executed as follows: 1) An additional feature, 'cv_label' was generated using the dataset's index (which in this case was AID, a unique identifier for each student in the dataset). This allowed us to specifically identify which instances were replicated during random oversampling. 2) RandomOverSampler from imblearn was applied to the dataset to balance the class counts. 3) Training and testing sets were then generated for the new dataset on an 80/20 split and 'cv_label' was separated into training and testing cross-fold labels and then removed from the training and testing sets. 4) The training

[10] There are numerous oversampling techniques such as SMOTE, which generates synthetic samples based on the k nearest neighbors of an instance. However, given the importance of cross-validation and the difficulty of establishing reliable cross-validation folds using SMOTE in the sklearn environment (fold identifiers are replicated as unique values using SMOTE) we opted for the simple approach of random oversampling.
[11] Documentation for imblearn can be viewed here: http://contrib.scikit-learn.org/imbalanced-learn/index.html

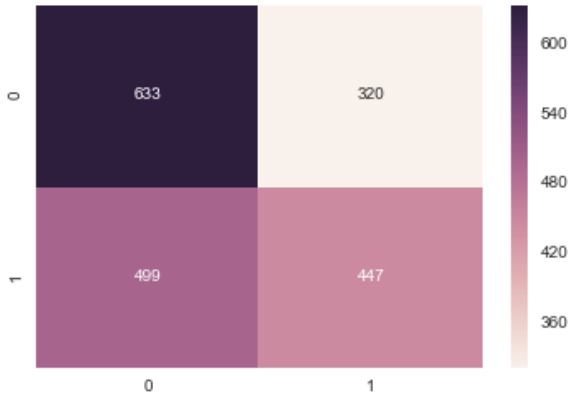
and testing cross-fold labels were then used to generate 10 validation folds for both the training and the testing set using LabelKFold from the sklearn.cross_validation module. 5) Subsequent predictions and parameter tuning were calculated using the generated training and testing folds to ensure that identical instances were not duplicated in a single validation fold. In the case of predictions, this was accomplished using cross_val_predict from the sklearn.cross_validation module with the cross-validation parameter set to the cross-validation labels for the testing set. In the case of parameter tuning (GridSearchCV) the cross-validation parameter was set to the cross-validation labels for the training set.

This technique was applied to both decision trees and random forest classifiers for each of the biomarker targets.

8a. Glucose Oversampling Decision Tree Model

Cross-validated predictions on the oversampled Glucose dataset provided significant improvements with a recall score of 0.47 and precision of 0.56 (entropy, max_depth=6, min_samples_leaf=61, min_samples_split=2). As expected, the precision and recall for the majority class were adversely affected, however, our focus is on knowledge discovery related to the minority class and the tradeoff was deemed acceptable.

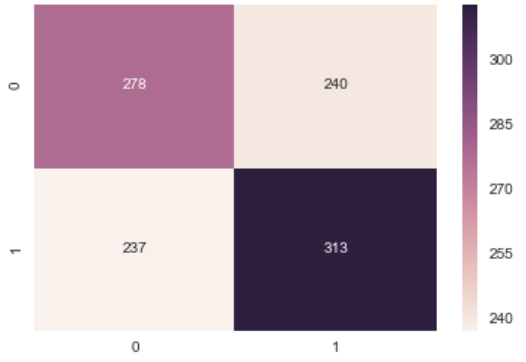
	precision	recall	f1-score	support
0	0.56	0.66	0.61	953
1	0.58	0.47	0.52	946
avg / total	0.57	0.57	0.56	1899



8b. Inflammation and Immune Function Oversampling Decision Tree Model (2 classes)

Utilizing the oversampling method on the 2-class variation of the Inflammation biomarkers produced a decision tree with 0.57 recall on the minority class with 0.57 precision (entropy, max_depth=10, min_samples_leaf=1, min_samples_split=22).

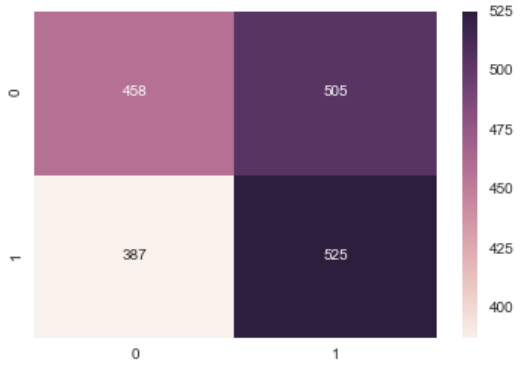
	precision	recall	f1-score	support
1.0	0.54	0.54	0.54	518
3.0	0.57	0.57	0.57	550
avg / total	0.55	0.55	0.55	1068



8c. Lipids Oversampling Decision Tree Model

Oversampling on the lipids dataset also provided significant improvements with a recall score of 0.58 on the minority class and a precision of 0.51 (entropy, max_depth=6, min_samples_leaf=11, min_samples_split=92).

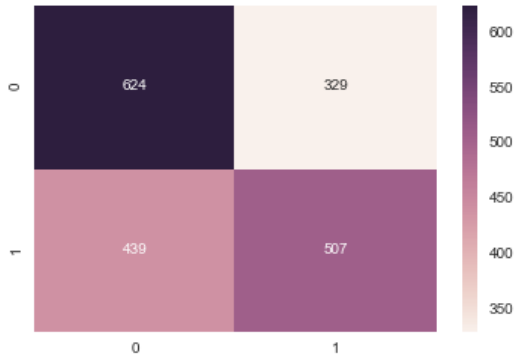
	precision	recall	f1-score	support
0	0.54	0.48	0.51	963
1	0.51	0.58	0.54	912
avg / total	0.53	0.52	0.52	1875



8d. Glucose Oversampling Random Forest Model

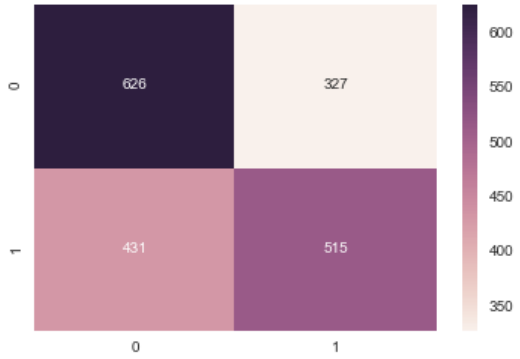
There are two models worthy of comparison here. The first allows the trees to grow to only a depth of one based on the gini criterion, with a min_samples_leaf=1, min_samples_split=82 and n_estimators=30. Its best result produces a recall of 0.54 for the minority class with a precision of 0.61.

	precision	recall	f1-score	support
0	0.59	0.65	0.62	953
1	0.61	0.54	0.57	946
avg / total	0.60	0.60	0.59	1899



A second model allows a max depth of 5 using the entropy criterion, with a min_samples_leaf=91, min_samples_split=72 and n_estimators=9. It can produce slightly higher recall for the minority class at 0.54 with a similar precision score of 0.61

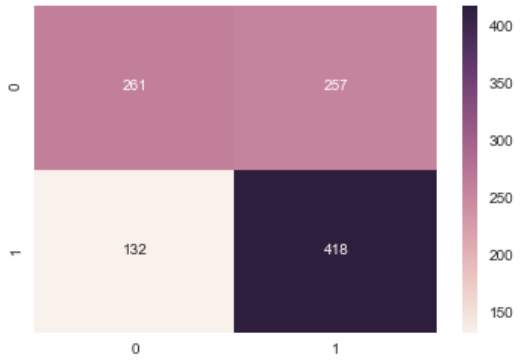
	precision	recall	f1-score	support
0	0.59	0.66	0.62	953
1	0.61	0.54	0.58	946
avg / total	0.60	0.60	0.60	1899



8e. Inflammation and Immune Function Random Forest Model

Utilizing the oversampling method on the 2-class variation of the Inflammation biomarkers produced a random forest with a recall score of 0.76 on the minority class with 0.62 precision (entropy, max_depth=5, min_samples_leaf=91, min_samples_split=72, n_estimators=90).

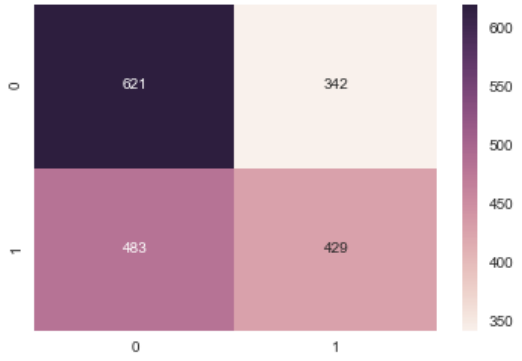
	precision	recall	f1-score	support
1.0	0.66	0.50	0.57	518
3.0	0.62	0.76	0.68	550
avg / total	0.64	0.64	0.63	1068



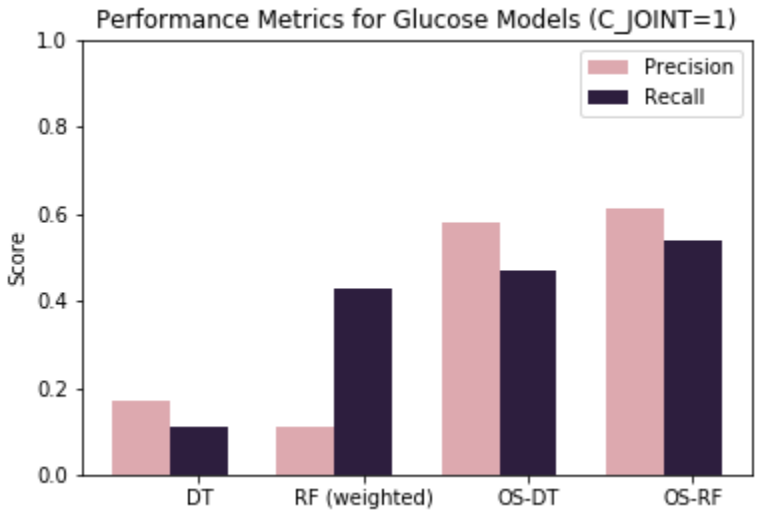
8f. Lipids Oversampling Random Forest Model

Interestingly, the best random forest model built on the oversampled dataset underperformed our best decision tree mode, perhaps indicative of the decision trees stability. The best generated recall score for the minority class was 0.47 with a precision of 0.56 (entropy, max_depth=6, min_samples_leaf-91, min_samples_split=72, n_estimators=100).

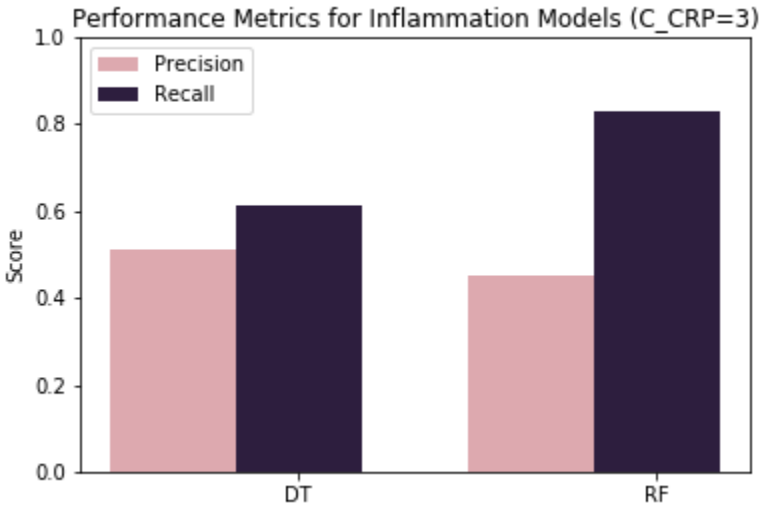
	precision	recall	f1-score	support
0	0.56	0.64	0.60	963
1	0.56	0.47	0.51	912
avg / total	0.56	0.56	0.56	1875



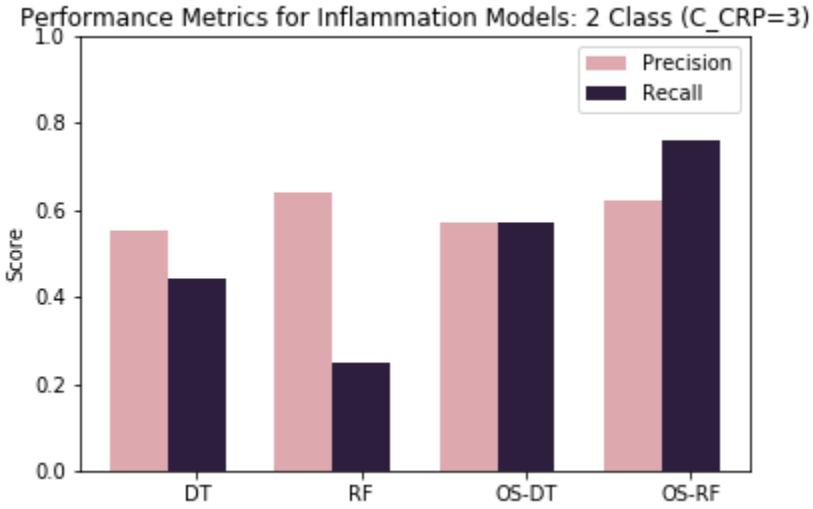
9. Comparison of Decision Tree and Random Forest models



A visual comparison of the models emphasizes how significantly random oversampling boosted our ability to capture information regarding the minority class. Both the oversampled decision tree and random forest (OS-DT, OS-RF) provided recall improvement along with significant precision improvement over the models built on the raw, imbalanced classes (DT, RF). The best overall performer was the oversampled random forest; however, the oversampled decision tree provides valuable support in interpreting the results.

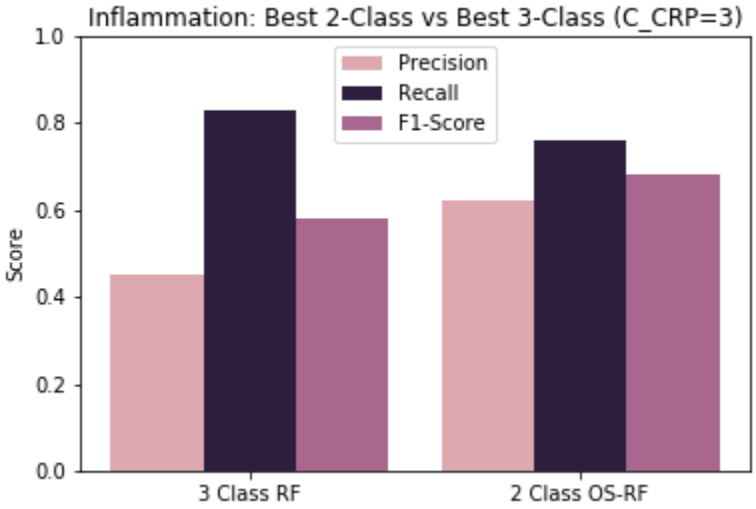


Models based on the original 3 classes for Inflammation performed reasonably well, with random forest providing the best overall performance.

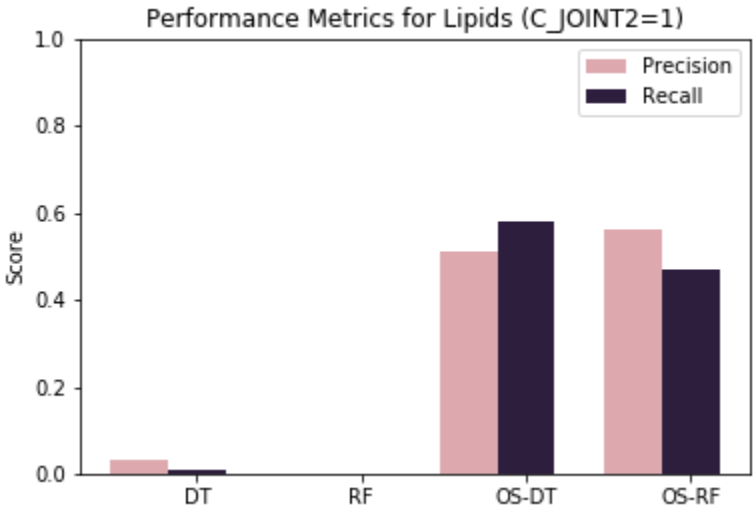


Initial models based on the 2-class version of C_CRP initially underperformed, however, unbalancing the classes allowed us to apply oversampling, which produced a decision tree and

random forest with comparable performance to the models based on the original 3 class structure.



If we compare the 3 Class Random Forest and 2 Class Oversampled Random Forest, we can see that the 2 Class variation outperformed the 3-class variation. While the 3-class model had higher recall, the 2-class model boosted precision with a minor decline in recall to generate a better f1-score.



The models for the lipids biomarker provide the most dramatic example of the value of oversampling with the oversampled decision tree and oversample random forest easily outperforming the original models, which were entirely lacking in utility.

10. Interpretation of Model Results

To review, the biomarkers were collected in 2008. The Wave 1 dataset collected information regarding these same students in 1994. Our models were attempting to capture information a full 14 years before the biomarkers were collected. Given the distance in time, we feel our models were able to capture a surprisingly large amount of information regarding the self-reported conditions of students in Wave 1. However, identifying patterns does not make them interesting or useful, so we will now turn to a more granular analysis of the model results. In the case of decision trees, the results are more straightforward to interpret. In the case of random forests, it is more difficult, as we cannot produce meaningful interpretation by looking at each of the individual estimators used to produce the results. In general, we will rely on the decision trees to guide analysis of the random forests when appropriate. Given the variety of features utilized in the model, we will focus on important features (features that clearly distinguished between classes) and interpretable (that is, for which a meaningful hypothesis can be generated). Visualizations of the trees used here are included as .png files with this report (we declined to include them as appendices due to resolution and sizing issues inside the document).

10a. Glucose

Interpretation of the glucose models come with some significant caveats. The dataset does not distinguish between Type 1 and Type 2 diabetes. The former is developed in childhood and is believed to be caused by a combination of genetic and environmental factors. The latter is developed later in life and is believed to be caused primarily by diet and is correlated with overweight individuals. However even Type 2 diabetes is believed to be partially genetically determined. Type 1 comprise approximately 5-10% of diabetes cases, while Type 2 accordingly comprises 90% of diabetes cases. We proceed tentatively with the assumption that most of the instances of diabetes in our dataset are Type 2 given the extent to which the dataset is nationally representative.

The tree classifies ~69% of nonwhite students as nondiabetic. 47% of non-white students classified as diabetic report that they are either “slightly overweight” or “overweight.” (H1GH28). Of these students, 80% of these students live in neighborhoods with median income of <= 34,500 per year (BST90P15). 80% of these students have mothers who do not have a college education are classified as diabetic (H1RM1). Of the 53% of non-white students classified as diabetic who do not report being overweight, 66% of them report watching more than 11.5 hours of television per week and 97% of these individuals have mothers who do not have a college education. A total of 1207 positive classifications of nonwhite diabetics (37% of all positive diabetic classifications) report having a mother with no college education.

Students who identify as white are split first based on the urbanicity of their neighborhood. The model is more likely to classify students living in "not completely urban areas" as diabetic (split of [405, 977] vs [753, 999] for those living in "completely urban" areas (BST90P01).[12]

Of those classified as diabetic in urban areas, 97% of them live in areas with a high dispersion of occupation type, however, we believe this is reflective of urban areas in general and not indicative of conditions that contribute to diabetes (BST90P25). Likewise, 86% of these individuals live in areas with balanced or heavily female sex composition (BST90P05). Of these individuals, 77% report that their mom is at home "always" or "most of the time" (H1RM11). Of those individuals who live in neighborhoods with heavily male composition all of those classified as diabetic (131 training instances in total) answered "yes" to the question, "Do your parents let you make your own decisions about the time you must be home on weekend nights?" (H1WP1). 68% of these individuals who self-report being overweight are classified as diabetic.

Among those classified as diabetic in "not completely urban areas," 89% of these individuals report that their people in their family understand them at least "quite a bit." (H1PR5). This is not particularly meaningful, however, of those individuals, 70% of those who live in areas with "medium" to "high" proportion of females aged 16 years and over in the civilian labor force are classified as diabetic (BST90P22). This represent 519 of 876 positive diabetic classifications in the parent node.

In general, a few trends emerge: Intuitively, whether an individual reports being overweight or not plays a significant role in our classifier. There is also a trend toward maternal involvement, with the mother's education level, presence in the house and the sex composition of neighborhoods playing crucial roles in classification. By extension, parental involvement is a frequent factor in classification and in one case is directly linked to self-reporting on weight status as the parent node. Median income plays a significant role in the case of nonwhite students, and this pattern would seem to be confirmed by our random forest, which leverages median income as its sixth most important feature (and the unemployment rate BST90P23 as its fourth most important feature). Our glucose decision tree is our weakest decision tree; however, it does manage to properly classify half of diabetic individuals with a precision of 0.58, and we feel comfortable suggesting that these areas merit, at the very least, special attention for future research.

10b. Inflammation and Immune Function

To review, C_CRP is a measure of inflammation and is useful in assessing for cardiovascular risk in asymptomatic individuals. Our interpretation will focus on the oversampling 2 class model. The decision tree can be viewed in appendices.

[12] "The urbanicity code distinguishes block groups that are in completely urbanized areas (BST90P01=1) from those that have any individuals living outside urbanized areas, in rural farm or rural nonfarm locations (BST90P01=2). " Wave 1: Public Contextual Database," p. 11.

The top feature for splitting is, which indicates whether an individual self-reports as overweight or not. 61% of training instances who report being "slightly" or "very" overweight are classified as having high hsCRP scores. The tree then splits on BST90P17 ("Median Family Income"). Individuals who live in neighborhoods with median family income of $17,500 per year or less represent only ~11% of the instances in the parent node. However, of the 167 training instances in this category, 79% of them are classified with high hsCRP scores, and of these 132 instances, 100% of the individuals classified as having high hsCRP scores report having never received a dental examination (H1GH25_4). Of the individuals living in areas with median family incomes of > $17,500, 88% of those classified with high hsCRP scores live in neighborhoods with "heavily male" sex composition (BST90P05). Of those 716 instances, 100% of them in which the resident father "went to a business, trade or vocational school instead of high school" (H1RF1_3) are classified as having high hsCRP scores.

Of those who do not report weight issues 98% of individuals who live in neighborhoods with occupation dispersion greater than 0.7065 are classified as having high hsCRP scores. While this is difficult to interpret meaningfully, the next split adds interesting depth. Of the 1156 high hsCRP classifications in this node, 1137 of those who report having used illegal drugs 4 or more times are classified as having high hsCRP.

With a recall score of .55 and precision of 0.6, we can be sure we are capturing at least some measure of meaningful information here beyond random chance. Self-reporting of weight plays an important role as with glucose levels. Median family income again appears, though this time it paints a clearer picture of the health risks for those who are low income. Perhaps most intriguing is the association between high hsCRP and frequency of dental examinations. While this split is made on only 132 instances in the training model, it does link intuitively to tentative studies linking heart disease to oral hygiene. As the Mayo Clinic indicates, there is preliminary evidence linking dental care to several diseases, including heart disease.[13] It is possible that oversampling amplifies a local health pattern that was previously suppressed by the imbalanced classes.

While our random forest model tends to focus on variables that indicate the physical maturity level of respondents, it does confirm some of the analysis, with H1GH28 as the top feature. Income also features as important, though median household income plays a more powerful role here than family income (BST90P15).

10c. Lipids

To review, C_JOINT2 indicates whether an individual is hyperlipidemic or not. Hyperlipidemia is abnormally elevated levels of lipids or lipoproteins and is viewed as a risk factor for

[13] http://www.mayoclinic.org/healthy-lifestyle/adult-health/expert-answers/heart-disease-prevention/faq-20057986

cardiovascular disease, pancreatitis etc. As with glucose, hyperlipidemia is probably partially genetically determined, so the results should be treated with care.

While the metrics for lipids are on par with our other models, the decision tree provides the least intuitive and interpretable results.

The top splitting criterion is H1GH28, which as noted above is a self-reported metric for weight. Of the 496 individuals who report that they are "very overweight," ~72% of them are classified as hyperlipidemic. Among these individuals, ~79% those who have smoked cigarettes before the age of 8 (H1TO2) are classified as hyperlipidemic. The split sample is small, but interestingly 57 of the 58 individuals whose mother works as a "sales worker, such as insurance agent, store clerk" are classified as hyperlipidemic. However, the low entropy score (0.1257) indicates we should not read into this result too much as this is a relatively expected result.

Beyond this it is difficult to gain traction when interpreting. For instance, of those who do not report being "very overweight," 96.8% of individuals classified as hyperlipidemic report watching less than 13.5 hours of video a week. Intuitively, we might expect higher time spent watching videos to be associated with a hyperlipidemic classification. Of these individuals 74% of the hyperlipidemic classifications are represented by those who work less than 39.5 hours of work a week during the summer. Of the 1457 individuals at this node who report working 40 hours or more, there is a slight trend toward hyperlipidemia (split [605, 852]) and 593 of the instances at this node are classified as hyperlipidemic if they live in a neighborhood with a high dispersion of marital status. Our suspicion in this case is that our tree is modeling the noise in the data rather than real information. Given the above-mentioned extent to which genetics play a role in hyperlipidemia, we would suggest that the only meaningful split in this tree is H1GH28, with weight being a known factor in hyperlipidemia.

The results for the random forest generally confirm this analysis, though it is worth noting that median family income (BST90P17) and dispersion in educational attainment (BST0P21) both appear as one of ten most important features. However, their lack of presence in the decision tree makes their role difficult to interpret.

10d. Overview

Two overarching themes unites the models, that of self-reporting on weight and median income levels. Individuals who report being overweight are consistently classified as diabetic and/or at risk for cardiovascular disease. While this is not revelatory, the presence of this feature in all three models is indicative that they are capturing meaningful data for each of the biomarkers. In addition, lower income was consistently aligned with a positive classification for a health risk. While this is a less obvious observation, socioeconomic status is heavily determinative of diet and by extension determinative of long term health outcomes.

For glucose levels, we must be careful interpreting the results as we do not have a clear mechanism to distinguish between type 1 and type 2, however, given the prevalence of type 2

diabetes, we can cautiously assume that much of the captured data is reflective of individuals with this specific type. In addition to weight, median household income played an important classification roles pointing toward the importance of socioeconomic status for general health. Additionally, maternal involvement and education level seemed to play a secondary role. Parental oversight and quantity of television watching also played minor roles. The latter two are more difficult to confidently interpret, but the general suggestion seems to be that students whose parents are better educated and more likely to monitor their activity have slightly better health outcomes when it comes to developing diabetes. These should not be received as absolute declarations, rather suggestions for deeper analysis to validate them as concrete patterns.

For hsCRP scores, at least one interesting pattern emerged beyond the role of weight. For a small subset of student instances, access to dental care was predictive of high hsCRP scores and, by extension, cardiovascular risk. It is possible that oversampling amplified a small, previously hidden pattern in the data. We feel it is worthy of further exploration, though this data set does not likely to provide the proper context in which to do so. Among those not reporting weight issues, usage of illegal drugs was associated with a high hsCRP score. While illegal drug use might certainly lead to higher inflammation levels, it is important to remember that the self-reporting predates the biomarkers by 14 years. More likely is that this pattern captures a behavioral pattern that leads to long term health risks up to and including adverse risk for heart disease.

Beyond the role of self-reporting on weight, our lipid models produced little by way of useful results despite reasonable performance metrics. One possibility is that the roles of genetics as a mechanism for hyperlipidemia simply makes it impossible to observe meaningful patterns beyond factors such as weight, which is a direct and known contributor. The results of the random forest model do indicate that median income and educational attainment play roles, but it is difficult to directly interpret to what extent and in what manner they do so.

11. KNN

After exploring and achieving acceptable performance from the random forest models, and dramatically improving the performance of the decision tree models with the use of oversampling, we approached the problem next with the k-nearest neighbors (KNN) algorithm. Considering the heavy imbalance for two of three dependent variables we were exploring, we approached the use of KNN with caution. As expected, due to the class imbalance, we did not achieve meaningful performance with the KNN classifier for the C_JOINT and C_JOINT2 dependent variables. C_JOINT is a joint bio classification for high blood A1C levels and self-reported use of anti-diabetic medication use. C_JOINT2 is a joint bio classification for self-reported history of hyperlipidemia and antihyperlipidemic medication use. Table 1 lists the final values for each class after the dataset had been pre-pre-processed.

Table 1

C_JOINT		**C_JOINT2**	
(0) No evidence of Diabetes	4,746	(0) No evidence of hyperlipidemia:	4,686
(1) Evidence of Diabetes	361	(1) Evidence of hyperlipidemia	421

We were working with a more balanced class distribution for the third dependent variable as presented in Table 2 and as expected we achieved better results with the KNN classifier. C_CRP is a bio classification for High Sensitivity C-Reactive Protein and can be used as an indicator of Cardio Vascular Disease risk.

Table 2

C_CRP 3-class		**C_CRP 2-class**	
hsCRP < 1 mg/L - Low	1,371	hsCRP low-average	2,669
hsCRP 1 - 3 mg/L - Average	1,289	hsCRP > 3 mg/L - High	1,879
hsCRP > 3 mg/L - High	1,870		

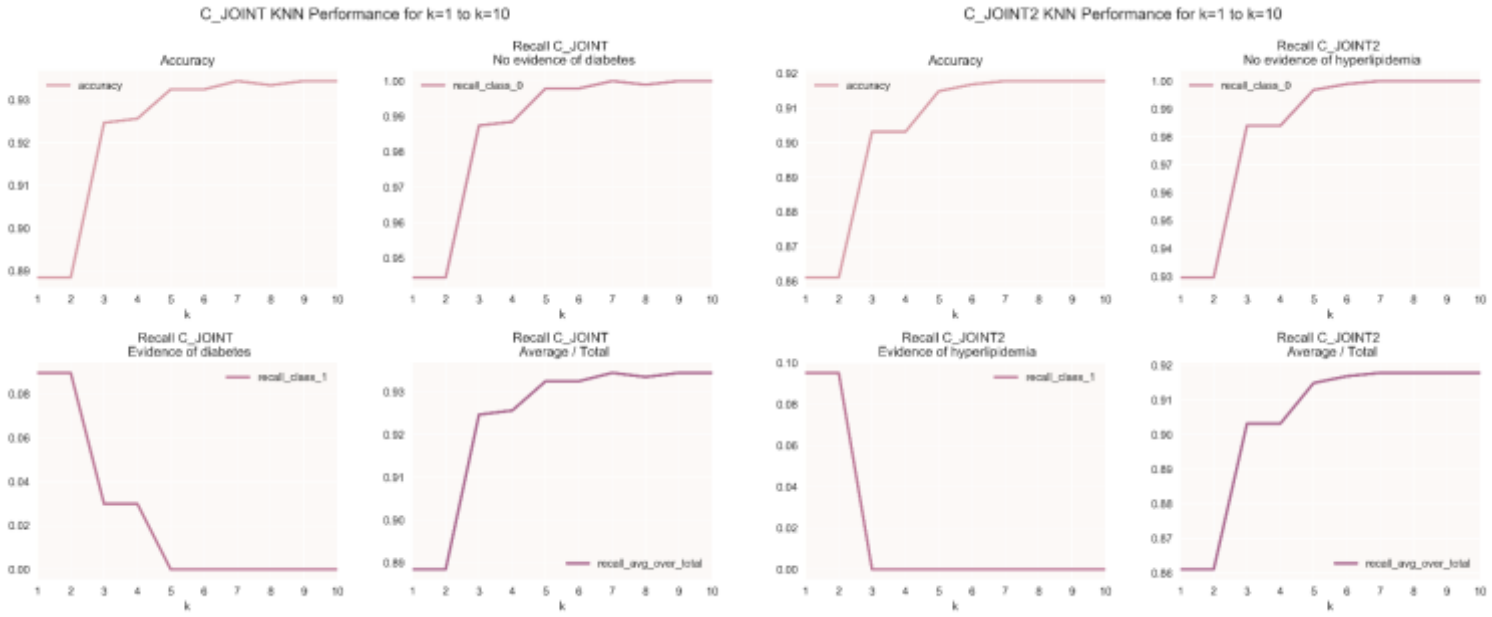
A similar approach was applied for each dependent variable in building the KNN classifier and can be examined in detail in the following Jupyter Notebooks, knn_glucose, knn_lip, and knn_immune (see repositories in appendix). The pre-processed dataset with appropriate dummy variables was loaded and examined. Target dependent variables were isolated and removed from the dataset. For the initial evaluation of k, the dataset was divided into an 80/20 train/test split. The metrics we identified most effective for evaluating the models were precision, recall, and accuracy, with recall being the most important metric. To determine the best value for k, an iterative function calculated the metrics of interest for a range of k=1 to k=100 depending on the dependent variable being examined. The KNN performance metrics were then plotted to help identify the ideal value range for k. After the value for k had been identified for each model, the classifier was further evaluated by running it against the entire dataset using 10-fold cross validation, rather than the initial 80/20 train/test split dataset.

The dataset was normalized to a range of 0 – 1 to account for the mix of categorical and numeric data. Normalizing to that range naturally leads to the use of the Euclidian distance metric. The weights parameter produced slightly stronger models when set to distance over uniform. Efforts were made to choose an odd value for k to avoid tie voting situations, however

due to the poor performance of the KNN classifier for the C_JOINT and C_JOINT2 dependent variables, an even value was chosen for C_JOINT2 out of necessity as the next odd value produced a non-predictive model for the key recall metric.

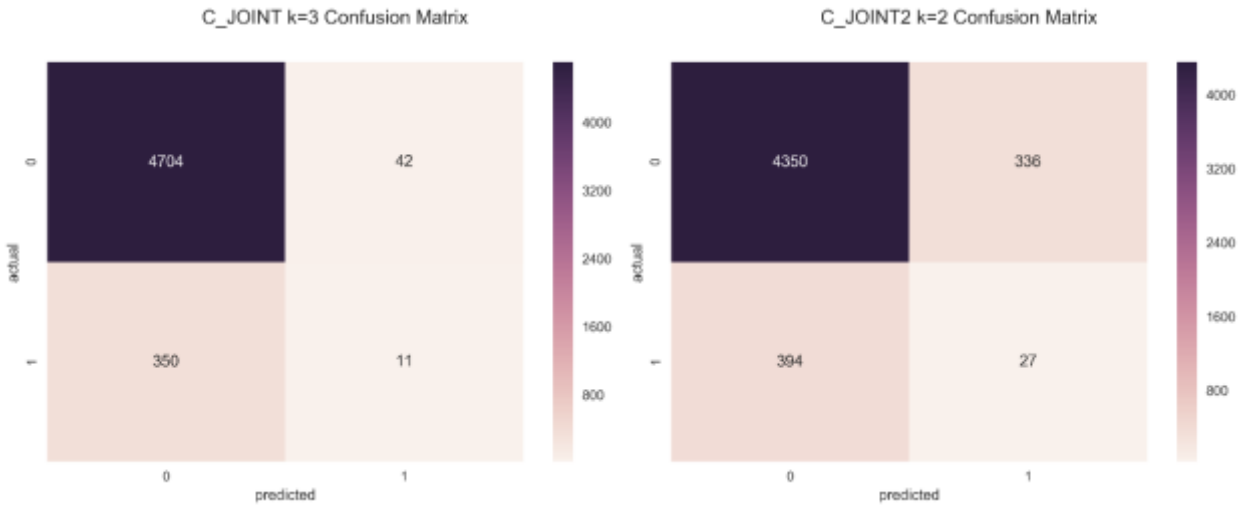
Due to the high class imbalance for C_JOINT and C_JOINT 2, very high levels of accuracy were obtained from the KNN models. As shown below, a high level of accuracy was achieved with a low value for k. However, this metric should be considered meaningless as the models simply predicted a negative class for diabetes or hyperlipidemia for all instances. The more important recall metric quickly fell to non-predictive levels rendering the models useless.

Figure 1



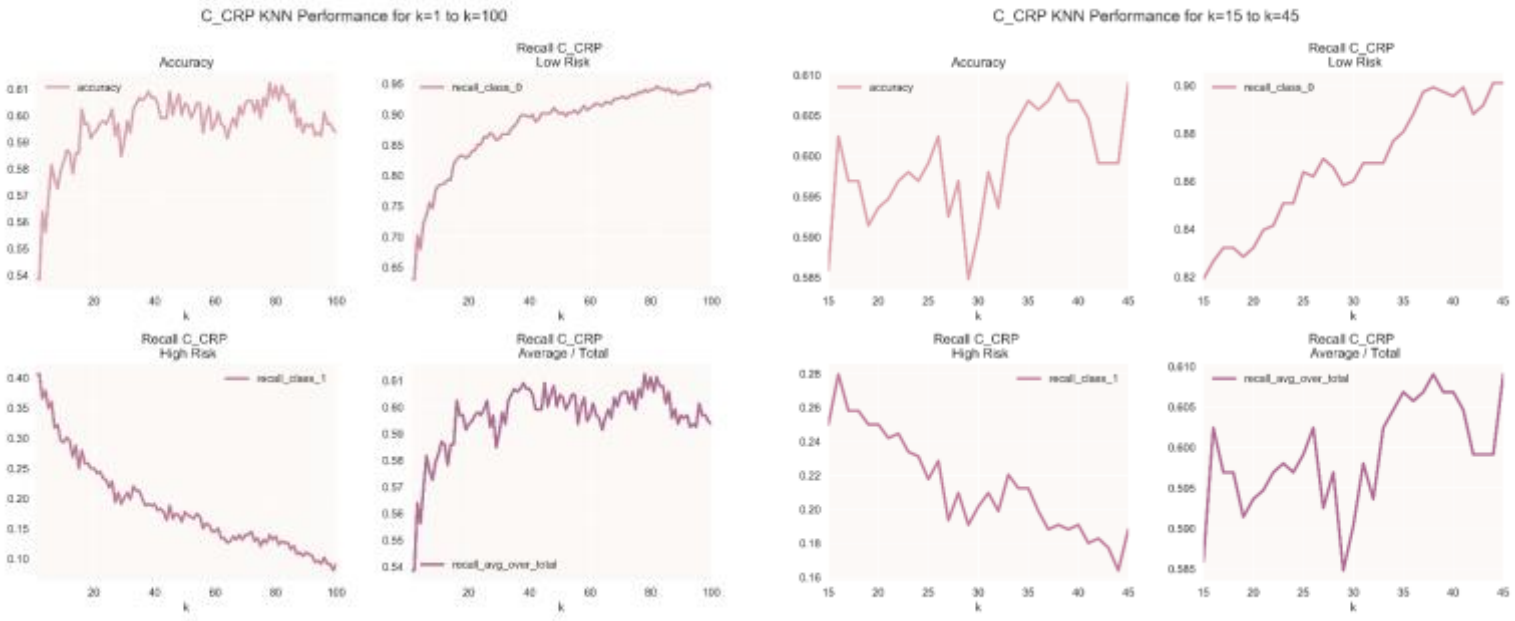
After examining Figure 2, we chose k values for final model evaluation of 3, and 2, for C_JOINT, and C_JOINT2, respectively. It could be argued that k should be set to 2 for C_JOINT resulting in a noted increase in recall, but that would come at the cost of an even value for k, in addition to a marked reduction in precision and accuracy. We can already see that the KNN model is not performing well with the imbalanced classes. The final model summaries after running 10-fold cross validation are presented in Figure 2.

Figure 2



Overall accuracy after 10-fold cross validation for C_JOINT and C_JOINT2 was 92 and 86 percent respectively. As we have already discussed, in this case, accuracy is a meaningless metric for model evaluation. We can see in the above matrices that the C_JOINT model misclassified 350 yes instances as no and C_JOINT2 misclassified 394 yes instances as no. This should be taken as a key indicator that the models are not performing and should not be considered for further evaluation especially when considering the high rate of Type II errors.

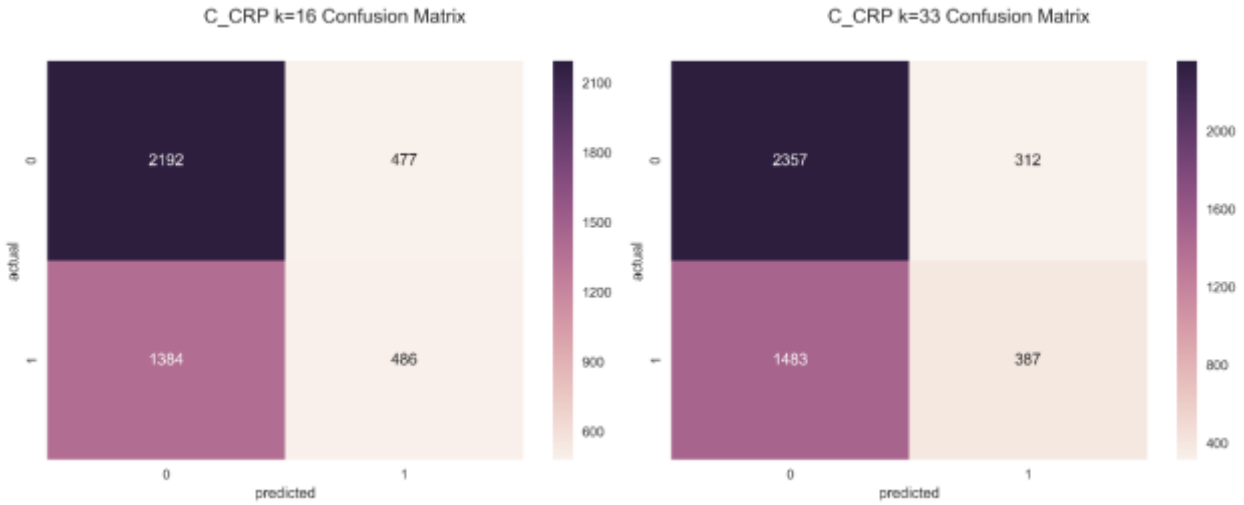
The application of the KNN algorithm to our dependent variables did not prove to be fruitless as we achieved significantly better results with the C_CRP dependent variable. As mentioned previously, we collapsed the classification for C_CRP from three classes to two classes, to define average vs high risk of developing Cardio Vascular Disease. The classes were more evenly



balanced leading to stronger KNN performance. An initial iterative run of k=1 to k=100 was performed as shown below. It is clear to see recall fall steadily as the value of k increases. We observed two peaks in both recall and accuracy in the range of k=15 to k=45, which called for closer examination.

All of our key performance measures peak at k=16. As the value of k increases, recall naturally falls, however, we observed another interesting peak at k=33 thus prompting us to examine both k=16 and k=33 for the final model.

Figure 4



After cross validation, it is clear to see the preferred model uses a value of k=16. Both models performed similarly when measured by accuracy with k=16 achieving an estimated accuracy of 59 percent and k=33 achieving an estimated accuracy of 60 percent. It is in the metric of most importance, recall and false negatives, where the k=16 model pulls ahead, misclassifying 1384 yes instances as no compared to k=33 misclassifying 1483 yes instances as no. Considering the similar performance for each model, we ultimately chose the k=16 model as it suffered from fewer Type II errors. Finally we briefly explored a k=17 model, resulting in a performance reduction that could not be justified, over the risk of tie votes with an even k value of 16. The complete classification report for each model is presented in Table 3.

Table 3

k=16					**k=33**				
class	precision	recall	f1-score	support	**class**	precision	recall	f1-score	support
1	0.61	0.82	0.7	2669	**1**	0.61	0.88	0.72	2669
3	0.5	0.26	0.34	1870	**3**	0.55	0.21	0.3	1870
avg/tot	0.57	0.59	0.55	4539	**avg/tot**	0.59	0.6	0.55	4539

12. Wave 4 Glucose Preliminary Results

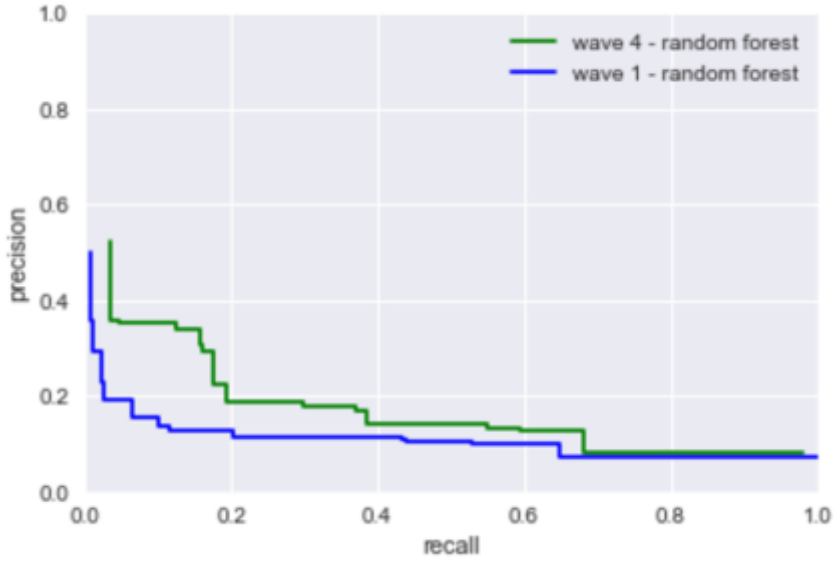
Given the difficulty of predicting diabetes based on Wave 1, we were interested in whether data in closer time proximity to the glucose biomarker would have higher predictive ability. So, we ran a quick analysis into the Wave 4 survey data to see if we could improve our results. Since Wave 4 survey questions were coincident to the biomarkers measurement (and diabetes diagnosis), we were no longer building a forecasting model. Instead, we were building a quantitative inference (diabetes 'diagnosis') model based on qualitative data at a common point in time. Admittedly, the value of such a model is lower than that of a forecasting model, but could still pay off by confidently identifying low and high-risk individuals to optimize the usage of expensive biomarker measurement for a subset of the population.

Moving our analysis to another wave of data presented a scalability challenge. We manually selected important features from Wave 1 with a divide and conquer approach in which we examined the entire dataset, but we did not have time to do the same for Wave 4. There are 27 sections with a total of 919 questions to choose between. Furthermore, we manually distinguished categorical and numeric questions in Wave 1 (to tell which ones to dummy), another thing we did not have time for in Wave 4.

We applied heuristics to solve each of these problems. To distinguish questions that needed to be dummied, we counted unique responses for each question and treated questions with 10 or fewer unique responses as categorical (and dummied them). With this treatment applied, our 919 questions resulted in 2,649 independent variables. Considering the hardware we had available, we did not think training on this full set was a good approach. We managed the size of our predictors by measuring the correlation between each and our target feature, and then selected the top 500 to arrive at a data set roughly the same size as we used in Wave 1 analysis.

Initial analysis showed a very promising predictor (H4ID5D), with a .63 Pearson correlation to our target. However, the feature represented a survey question about whether the respondent had ever been diagnosed with diabetes. After excluding H4ID5D, we moved into modeling the data. We applied GridSearchCV to explore the precision recall tradeoff with Wave 4 data, but time restrictions prohibited applying oversampling techniques in Wave 4.

As expected, we did find higher predictive ability with Wave 4 data. In cross validation, our preferred model had an average recall of .55 and average precision of .14, slightly better than the comparable model in Wave 1 (recall = .43, precision = .11). We plotted the best random forest models along the precision-recall tradeoff from Wave 1 and Wave 4 to show the improvement we see with Wave 4 data.



The initial improvement is marginal, but gives further credibility to our assertion that predicting diabetes from the data at hand (especially without distinction between Type I and Type II) is a difficult task hampered by our inability to distinguish between diabetes type as well as account for potential genetic inheritance. We are eager to extend additional techniques to the Wave IV questionnaire as well as consider the other biomarkers in this context.

Appendix

This appendix contains links to repositories for reports, complete scripts and third-party packages used in this report.

Initial Survey of Wave 1 Variables

https://drive.google.com/open?id=1dZvSiee16-cm4O0k5aP2zeHQr6LM-9B0CLyGKM_-OYw

Data Preparation Scripts and Cleaned Datasets

https://drive.google.com/open?id=0BzbnkL9_iTY5QTRWZVhwUy1Tdlk

Model Scripts

https://drive.google.com/open?id=0BzbnkL9_iTY5ZlRKVzdRQlJuUGs

Visualization Scripts and Visualization Graphics

https://drive.google.com/open?id=0BzbnkL9_iTY5OXpyRGhGNEtSZFE

Introductory Information and Documentation for imblearn (used for oversampling technique)

http://contrib.scikit-learn.org/imbalanced-learn/index.html